Phase 1: EDA

You will be provided with a dataset including over 5 million tweets in a certain time period mentioning at least one of the stocks traded in the US markets as well as primitive information about the users making those tweets. You will also be given financial information about those stocks. The task is to first perform a thorough explanatory data analysis to gain a brief and holistic vision on the data you have at hand and what you can do with it. Creative visualization methods are rewarded!

Your results should include creative and explanatory plots and graphs on <u>at least</u>:

- Statistics on most (least) tweeted stocks
- Statistics on distributions of tweets including individual stocks (at least 2) over time
- Statistics on distributions of all financial tweets over time
- Statistics on distribution of retweets per tweets including individual stocks (at least 2)
- Statistics on most important financial information on individual stocks (at least 2) computed solely from the financial information (not the tweets)
- Time series movement directions through time for individual stocks (at least 2)
- (Optional) co-occurrence of various stocks in the same tweets